



ARTICLE

A Transformer-Based Multimodal Cross-Attention Framework with Relational Graph Attention for Pill-Prescription Matching

Linh Nguyen Thi My^{1,2,*}, Lap Thai Viet², Hoai Truong Hieu², Tham Vo¹ and Vinh Truong Hoang^{3,*}

¹Faculty of Information Technology, Nguyen Tat Thanh University, 300A Nguyen Tat Thanh Street, Xom Chieu Ward, Ho Chi Minh City, Vietnam

²Faculty of Information Technology, Van Lang School of Technology, Van Lang University, 69/68 Dang Thuy Tram Street, Binh Loi Trung Ward, Ho Chi Minh City, Vietnam

³Faculty of Information Technology, Ho Chi Minh City Open University, 35–37 Ho Hao Hon Street, Cau Ong Lanh Ward, Ho Chi Minh City, Vietnam

*Corresponding Authors: Linh Nguyen Thi My. Email: 2311561010@nttu.edu.vn; Vinh Truong Hoang. Email: vinh.th@ou.edu.vn

Received: 06 June 2026; Accepted: 19 August 2026; Published: 22 October 2026

ABSTRACT: Matching physical pills with drug names on a prescription is a safety-critical task for preventing medication errors. The pioneering PIMA (Pill-prescription MAtching) framework addressed this problem by aligning convolutional features (ResNet50) with textual embeddings (BERT) using a direction-agnostic GraphSAGE operator and a margin-based contrastive loss. While effective on controlled datasets, PIMA's real-world deployment is hindered by its reliance on manual ground-truth annotations, sensitivity to background clutter, and a simplistic cosine similarity fusion that lacks selective attention. In this paper, we propose a redesigned, fully multimodal framework featuring five coordinated improvements: (i) two interchangeable visual encoders for the pill-recognition branch—a Vision Transformer (ViT-B/16) applied to background-removed crops, and an end-to-end Faster R-CNN/RoI-Align detector that learns features directly from raw images; (ii) a lightweight all-MiniLM-L6-v2 sentence encoder for efficient pill-name representation; (iii) an integrated PP-OCRv3 module that automates text and bounding box extraction, eliminating manual dependency; (iv) a Relational Graph Attention Network (R-GAT) with a pseudo-classifier gate to model directional spatial logic (e.g., left-to-right drug/quantity relations) and filter irrelevant non-drug text; and (v) a multimodal cross-attention head optimized via an InfoNCE objective for more stable and dense gradient convergence. Experiments on the real-world VAIPE dataset show that both variants substantially outperform a reproduced graph baseline: the ViT-B/16 variant raises the Top-1 matching accuracy from 49.89% to 82.46% and the end-to-end Faster R-CNN variant to 83.47%—absolute gains of 32.57 and 33.58 percentage points (pp), respectively, while placing the correct drug name among the top-five candidates for more than 99% of pills. The two variants trade off accuracy against preprocessing: the Faster R-CNN variant attains the best accuracy while operating directly on raw images, whereas the ViT-B/16 variant requires a separate background-removal step. In terms of computational footprint, the visual backbones contain approximately 86.6M (ViT-B/16) and 41.8M (Faster R-CNN/ResNet50-FPN) parameters, the 22.7M-parameter MiniLM encoder replaces the far heavier BERT of the baseline, and although training was distributed over four GPUs, inference requires only a single commodity GPU. Our system also remains robust to optical-character-recognition (OCR) noise, showing minimal accuracy degradation when switching from ground-truth to automated text extraction. These results confirm that a Transformer-based cross-attention architecture with directional spatial awareness provides an effective and deployable solution for automated pill-prescription verification.

KEYWORDS: Pill-prescription matching; multimodal learning; vision transformer; relational graph attention; cross-attention; InfoNCE; medication safety

1 Introduction

Medication errors represent a persistent and global threat to patient safety [1]. According to the World Health Organization (WHO), avoidable harm in healthcare is frequently attributable to the administration of incorrect drugs, wrong dosages, or non-prescribed medications [2]. Published estimates of the resulting mortality vary widely with the definition of error and with the reporting system used, so no single figure can be regarded as authoritative; what is consistent across sources is that medication errors remain a leading cause of avoidable and largely preventable harm in health care worldwide [2]. Recognizing the severity of this issue, the WHO designated “Medication Without Harm” as the theme for World Patient Safety Day 2022 [2]. Automated verification systems that can validate medications from ordinary mobile photographs against written prescriptions therefore hold immense clinical value [1,3]. Consequently, pill image recognition has transitioned from a niche task to an active research area spanning computer vision, optical character recognition (OCR), and clinical informatics [1].

More broadly, deep learning has become the enabling technology of modern computer-aided healthcare because it learns discriminative representations directly from raw clinical data, removing the dependence on brittle hand-crafted features and enabling decision support at the speed demanded by clinical workflows. In medical imaging, deep neural networks now approach or exceed expert-level performance on diagnostic tasks; for example, the M3BTCNet framework couples multi-model CNN features with metaheuristic feature optimization to achieve reliable multimodal brain-tumor classification [4]. In the real-time healthcare sector in particular, deep models embedded in mobile or edge devices can verify medications at the bedside or in a patient’s home within fractions of a second—a level of speed, consistency, and scalability that manual double-checking cannot offer [5,6]. Pill-prescription verification is a natural target for this technology: the task is repetitive, vision-intensive, and safety-critical, which is precisely the regime in which automated deep learning support is most valuable.

While deep learning models have revolutionized isolated pill classification, real-world verification requires more than identifying a single pill. A recent comprehensive survey [2] highlights that the most realistic and challenging formulation is not classification in a controlled setting, but the joint reasoning required to connect a photograph of multiple pills to the structured-but-noisy text of a prescription. This task, formalized as pill-prescription matching, requires a model to bridge the semantic gap between visual pill morphology and textual drug nomenclature [3]. Unlike general text-image matching, drug names often lack inherent semantic meaning, are lengthy, and are embedded within a layout filled with irrelevant information such as dosages and diagnoses [3,7]. This line of work originates from a sustained research effort on Vietnamese real-world pill data: contextual multi-pill recognition aided by a medical knowledge graph (PIKA) [8], class-incremental pill classification [9], and graph-assisted multi-pill detection on the open VAIPE dataset [10]. Within this effort, the PIMA (Pill-prescription MAtching) framework [3] was the first to define and address the pill-prescription matching task using a graph-based representation of prescriptions, and it was later extended to a zero-shot setting (Zero-PIMA) [7].

PIMA [3] modeled prescriptions as graphs whose nodes represent text boxes and whose edges capture spatial proximity, using a GraphSAGE operator to identify drug names. Its architecture, however, has four bottlenecks that limit real-world deployment. It relies on a convolutional backbone (ResNet50) [11] that is sensitive to background clutter and struggles to capture fine-grained morphological cues. It also assumes access to manual ground-truth annotations for text and bounding boxes, which makes it non-functional for end-to-end use without human intervention. In addition, it aggregates spatial context with a direction-agnostic operator that cannot distinguish directional layout logic, such as a dosage quantity typically appearing to the right of a drug name. Lastly, it fuses the two modalities with a simple cosine similarity that lacks selective attention, preventing the model from focusing on the most relevant textual

evidence for a given pill's appearance. These limitations are not merely conceptual but measurable: the original study reports a matching accuracy of 46.95% [3], and our faithful reproduction of the full pipeline on the VAIPE dataset reaches only 49.89% (Section 5), meaning that roughly one in every two pills would be verified incorrectly—an unacceptable rate for a safety-critical clinical application. Section 5 further quantifies the individual bottlenecks through ablation, providing direct empirical evidence for the necessity of the redesign proposed here.

In this work, we argue that a Transformer-based multimodal design can jointly overcome these limitations. We propose a redesigned framework featuring: (i) a Vision Transformer (ViT-B/16) for global feature extraction and a Faster R-CNN/RoI-Align variant for end-to-end processing of raw images [12,13]; (ii) a compact all-MiniLM-L6-v2 sentence encoder for efficient textual representation; (iii) an integrated PP-OCrv3 reader to automate the entire text extraction pipeline [14]; (iv) a Relational Graph Attention Network (R-GAT) with a pseudo-classifier gate to model directional spatial relations and filter non-drug text; and (v) a multimodal cross-attention head optimized via InfoNCE for stable gradient convergence. Unlike previous pair-wise margin losses, InfoNCE contrasts positive pairs against all negative samples in a batch, providing a much denser learning signal.

The main contributions of this paper are summarized as follows: we identify and analyze the architectural bottlenecks of the original PIMA framework and propose a coherent set of five enhancements to address them; we design a fully multimodal architecture coupling Transformer-based encoders with an R-GAT spatial encoder and a cross-attention fusion head; we integrate automated OCR and demonstrate that our framework, evaluated on the real-world VAIPE dataset [10], raises Top-1 matching accuracy from 49.89% (for the reproduced graph baseline) to 82.46% (ViT-B/16) and 83.47% (Faster R-CNN); and we provide an ablation study confirming that our system remains robust to OCR noise, maintaining high performance even when processing raw, unprocessed photographs, thus providing a deployable solution for automated pill verification.

The remainder of this paper is organized as follows. Section 2 reviews related work on pill recognition, object detection, graph neural networks, and multimodal contrastive learning. Section 3 details the proposed framework and its differences from the baseline. Section 4 describes the dataset, implementation, and evaluation protocol. Section 5 reports and discusses the results, including training curves and an ablation study. Section 6 concludes the paper and outlines future work.

2 Related Work

Our work builds on four threads of prior research: pill image recognition, object detection and visual backbones, graph neural networks (GNNs) for structured document understanding, and multimodal contrastive learning. We follow the taxonomy established in the recent comprehensive survey of pill image recognition [2].

2.1 Pill Image Recognition

Early pill-identification systems relied on hand-crafted descriptors of the three canonical pill attributes—shape, color, and imprint [2]. Shape was commonly captured with edge detectors such as Canny [15] and with moment invariants [16]; imprint and texture were described with local features such as SIFT [17] and Local Binary Patterns (LBP) [18]. While interpretable, such pipelines are inherently brittle: their performance degrades sharply when the imprint is worn, the lighting changes, or the background is cluttered, requiring careful manual tuning for each new dataset [2].

The subsequent shift to deep learning replaced these manual descriptors with features learned end-to-end. Convolutional networks such as VGG [19] and ResNet [11] became standard backbones for pill classification and retrieval, and underpin dedicated pill benchmarks and inspection models such as the ePillID benchmark [20] and deep pill-detection systems [21]. To cope with the fine-grained, low-inter-class-variance nature of medications, multi-stream variants that fuse several views or modalities have been proposed, including class-incremental multi-stream fusion [9] and metric-based recognition that combines textual and visual cues [22]. The ePillID benchmark [20] established a key finding that fine-grained pill identification is essentially a low-shot problem—most appearance classes contain very few images—and that standard classifiers degrade severely in this regime, although its images are captured under controlled reference conditions rather than in the wild. Deep detection systems such as [21] demonstrated reliable pill localization in dispensing scenarios, but they stop at detection and do not connect pills to a prescription. The most recent systems push this direction further toward deployable, real-world pipelines: clinical multi-platform studies of code-free deep learning [23] show that automated model construction lowers the entry barrier for clinical users but offers limited architectural control; mobile medication identification [24] and real-time inspection frameworks [5] confirm the feasibility of fast on-device inference while leaving the multimodal pill–prescription matching problem unaddressed. Beyond pill imagery, M3BTCNet [4] couples multi-model deep features with metaheuristic feature optimization for multimodal brain-tumor classification, illustrating that careful feature selection and fusion—rather than ever-larger backbones—often drives accuracy gains in medical imaging; the same principle motivates the attention-based multimodal fusion adopted in this work. Closely related to our setting, a multimodal framework that fuses RGB, contour, texture, and text cues through an adaptive lightweight-attention mechanism has been shown to achieve accurate real-time pill recognition on resource-constrained edge devices [6], underscoring the value of attention-based multimodal fusion that we also adopt in this work. In parallel, the field has begun to adopt Transformer-based recognition [2]. This trend motivates our adoption of a Vision Transformer (ViT) backbone, which uses self-attention to capture global morphological cues that traditional convolutional neural networks (CNNs) often struggle to resolve in noisy or unconstrained environments.

2.2 From Single-Pill Recognition to Pill–Prescription Matching

Recognizing an isolated pill is insufficient for verifying a prescription: the system must also decide *which* prescribed drug each detected pill corresponds to. This problem has been driven largely by a single, sustained line of research on real-world Vietnamese pill data. The group first introduced contextual multi-pill recognition with a medical knowledge graph (PIKA) [8], then addressed catastrophic forgetting through a multi-stream class-incremental learning framework (CG-IMIF) [9], and tackled multi-pill localization with a graph-assisted, explainable detection framework that also released the large-scale, open VAIPE dataset used in this study [10].

Within this effort, Nguyen et al. defined the pill–prescription matching task together with the PIMA (Pill-prescription MAtching) framework [3]. PIMA builds a graph over the text boxes of a prescription, propagates information with GraphSAGE to emphasize boxes containing pill names, and matches the highlighted text embeddings against convolutional pill features with a contrastive loss, supported by an auxiliary classification loss. The framework was subsequently extended to recognize pills unseen during training in a zero-shot setting (Zero-PIMA), where a graph convolutional network and metadata-enriched name embeddings improve matching for novel drugs [7]. PIMA is the baseline that the present paper improves upon: we retain its problem formulation and its use of a text-box graph, but we replace its visual backbone, its graph operator, its language encoder, and its fusion mechanism, and we remove its dependence on ground-truth annotations.

2.3 Object Detection and Visual Backbones

Because pills must be localized within a photograph before they can be matched, object detection is central to an end-to-end system. Two-stage detectors such as Faster R-CNN [12], strengthened with feature pyramid networks [13] and RoI-Align region features [25], offer accurate localization, while one-stage detectors such as [26], SSD [27], and [28] trade some accuracy for speed. Segmentation networks such as U-Net [29] and Mask R-CNN [25] are frequently used to isolate pills from complex backgrounds [2]. In our end-to-end variant we adopt a Faster R-CNN/ResNet50-FPN detector and use RoI-Align to extract a feature vector for each pill directly from the raw image, avoiding a separate background-removal stage.

2.4 Text Extraction from Prescriptions

Reading the drug names from a prescription requires optical character recognition. Classical engines such as Tesseract [30] perform well on clean, printed text but struggle with the noisy, heterogeneous layouts of real prescriptions. Modern OCR systems based on convolutional–Transformer hybrids, in particular PP-OCrv3 [14], combine text detection and recognition and are far more robust under such conditions [2]. We embed PP-OCrv3 in our pipeline so that text content and bounding boxes are produced automatically, replacing the ground-truth annotations assumed by the baseline.

2.5 Graph Neural Networks and Multimodal Contrastive Learning

Graph neural networks are a natural tool for the spatial layout of a prescription, where the relevant signal is the relative position of text boxes. The baseline uses GraphSAGE, which aggregates neighbors without regard to their direction. Relational graph attention [31], by contrast, learns separate attention weights for typed (directional) edges, allowing the model to encode that a quantity usually appears to the right of a drug name or that a diagnosis appears above the medication list. For the cross-modal step, contrastive objectives that pull matched image–text pairs together while pushing apart mismatched pairs have become the de-facto standard, as demonstrated at scale by CLIP for image–text alignment [32]; the InfoNCE loss [33] in particular optimizes a softmax over all pairs in a batch and provides a stronger learning signal than a fixed cosine threshold. Our framework combines a relational graph attention encoder with a cross-attention fusion module and an InfoNCE objective, an instantiation of these ideas tailored to pill–prescription matching.

2.6 Summary and Research Gap

Table 1 summarizes the most closely related systems, their key findings, and the gaps that remain with respect to fully automated pill–prescription verification. In summary, no existing system simultaneously (i) operates end-to-end on raw photographs without manual text or bounding-box annotation, (ii) models the directional layout logic of prescriptions, and (iii) fuses the two modalities with selective attention trained by a dense contrastive objective. Closing this specific gap is the goal of the framework proposed in the next section.

Table 1: Representative related works, their key findings, and the gaps that remain for automated pill–prescription verification.

Work	Approach	Key Finding	Gap for Our Task
PIKA [8]	Knowledge-graph-assisted contextual pill recognition	Prescription-based knowledge graph boosts recognition of multiple pills in an intake photo	Classification only; pills are not matched against a given prescription

(Continued)

Table 1 (continued)

Work	Approach	Key Finding	Gap for Our Task
CG-IMIF [9]	Class-incremental multi-stream fusion	Mitigates catastrophic forgetting as new drugs are added	No text modality; classification only
VAIPE detection [10]	Graph-assisted multi-pill detection	Co-occurrence reasoning improves detection; released the open VAIPE data	Stops at localization; no matching
PIMA [3]	GraphSAGE + margin-based contrastive matching	First formalization of pill–prescription matching (46.95% reported accuracy)	Manual annotations; direction-agnostic graph; cosine fusion
Zero-PIMA [7]	GCN + metadata-enriched name embeddings	Matches drugs unseen during training	Inherits PIMA’s visual backbone and fusion
ePillID [20]	Low-shot fine-grained benchmark	Standard classifiers fail in low-shot pill identification	Controlled images; no prescriptions
Edge multimodal ALA [6]	Adaptive lightweight-attention fusion	Real-time multimodal recognition on edge devices	Single-pill recognition; no layout reasoning

3 Proposed Method

3.1 Overview and Problem Formulation

Given a prescription image P and an associated image I containing one or more pills, the goal of pill–prescription matching is to assign to every detected pill v the drug name on the prescription that it corresponds to, or to report that no match exists. Let the prescription contain M text boxes with textual content $\{t_1, \dots, t_M\}$ and spatial coordinates $\{b_1, \dots, b_M\}$, and let the pill image yield N pill regions $\{x_1, \dots, x_N\}$. The model must produce, for each pill x_i , a probability distribution over the candidate drug-name boxes.

The proposed framework, shown in Fig. 1, comprises three encoders and a fusion head. A visual encoder maps each pill region to a feature vector; a text encoder maps each prescription text box to a sentence embedding; a relational graph encoder refines these embeddings using the directional spatial layout of the prescription and predicts, through a pseudo-classifier, which boxes are likely to contain pill names; and a multimodal cross-attention head fuses the visual and textual representations and is trained with an InfoNCE contrastive objective. Unlike the baseline PIMA [3], the text content and bounding boxes are not assumed to be given: they are produced automatically by an integrated PP-OCRv3 reader [14].

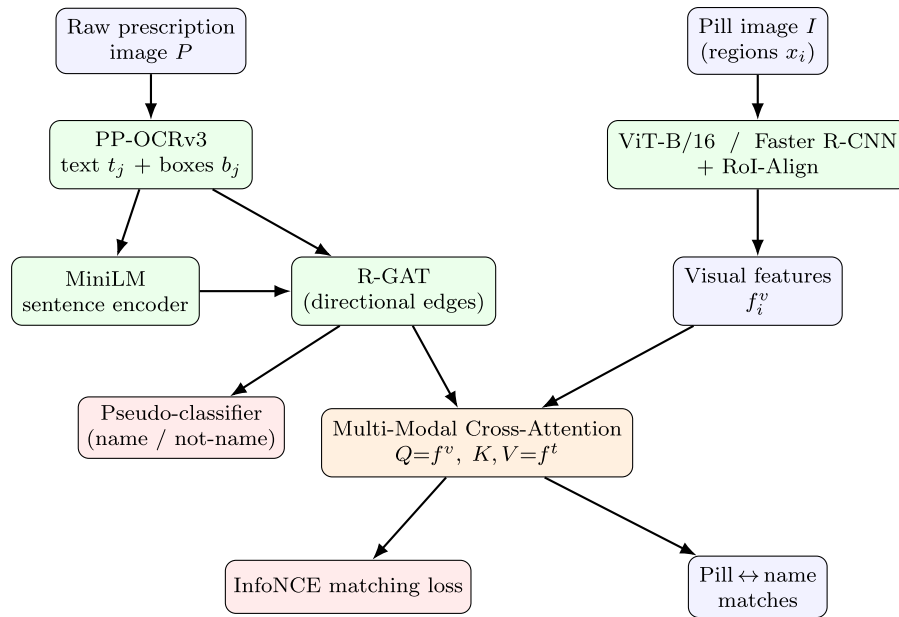


Figure 1: Overview of the proposed Transformer-based multimodal framework. Green blocks are learnable encoders, the orange block is the cross-attention fusion head, and red blocks are training objectives. The PP-OCRv3 reader removes the dependence on ground-truth text and bounding boxes assumed by the baseline.

3.2 Visual Encoder

The original PIMA framework extracts pill features with a ResNet50 convolutional backbone. We replace this with two interchangeable visual encoders: a Vision Transformer that models the pill as a sequence of image patches, and an end-to-end region-based convolutional detector.

ViT-B/16 (clean-crop variant). Each pill is first separated from its background (e.g., with a matting tool) and the clean crop is fed to a Vision Transformer ViT-B/16. By splitting the image into 16×16 patches and modeling them with self-attention, the ViT relates distant regions of the pill surface to one another and thereby captures global morphological cues—outline, color distribution, and imprint. The background is eliminated beforehand by the matting step rather than by the self-attention mechanism, so the quality of this preprocessing bounds the accuracy of the variant.

The core operation driving this architecture is the Scaled Dot-Product Attention:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \quad (1)$$

where Q , K , and V denote the query, key, and value matrices obtained through learnable linear projections of the patch-token embeddings, d_k is the dimensionality of the key vectors, and the softmax is applied row-wise. This mechanism allows the model to establish long-range dependencies across the entire pill image, ensuring that fine-grained details like imprints are not lost in background noise (which is a common limitation of the ResNet50 backbone). Finally, the class token produces a d -dimensional pill embedding f_i^v .

Faster R-CNN (end-to-end variant). To avoid a separate background-removal step, the second variant feeds the raw image, together with pill bounding boxes, to a Faster R-CNN with a ResNet50-FPN backbone [12,13]. RoI-Align [25] extracts a fixed-size feature map for each pill region, which is pooled into the

embedding f_i^v . This variant is fully end-to-end and therefore better suited to deployment on unprocessed photographs, at the cost of having to learn to suppress background noise.

3.3 Language Encoder and OCR

Drug names are extracted from the prescription by PP-OCRv3 [14], which jointly detects text regions and recognizes their content, yielding for each box j a string t_j and coordinates b_j . Each string is embedded by the compact all-MiniLM-L6-v2 sentence encoder [34,35], producing a textual feature $f_j^t \in \mathbb{R}^d$. Compared with the heavier BERT encoder used previously, MiniLM provides sentence embeddings at a fraction of the computational cost while preserving semantic quality, which is important because a prescription contains many boxes that must each be encoded.

3.4 Relational Graph Attention Encoder

The prescription is represented as a graph $G = (\mathcal{V}, \mathcal{E})$ whose nodes are the text boxes and whose edges encode directional spatial relations (left, right, above, below) derived from the boxes' coordinates. Whereas the direction-agnostic GraphSAGE aggregates neighbors identically regardless of direction, we use a Relational Graph Attention (R-GAT) operator. This explicitly models the directional spatial logic of a prescription (e.g., drug name $\xrightarrow{\text{right}}$ quantity) by learning a separate transformation and attention coefficient per relation type r :

$$h_j^{(l+1)} = \sigma \left(\sum_{r \in \mathcal{R}} \sum_{k \in \mathcal{N}_r(j)} \alpha_{rjk} W_r h_k^{(l)} \right), \quad (2)$$

where $\mathcal{R} = \{\text{left, right, above, below}\}$ is the set of directional relation types, $\mathcal{N}_r(j)$ is the set of neighbors of node j connected by relation r , W_r is a relation-specific learnable weight matrix, $h_k^{(l)}$ is the representation of node k at layer l , and $\sigma(\cdot)$ is a non-linear activation function. By using these distinct matrices for each direction, the model effectively reconstructs the logical structure of the document, which is critical for identifying which text boxes actually belong to a specific medication entry. The attention weight α_{rjk} is obtained by a softmax over a learnable compatibility score between $h_j^{(l)}$ and $h_k^{(l)}$. The node features are initialised with the MiniLM embeddings, $h_j^{(0)} = f_j^t$.

A *pseudo-classifier* gate is attached to the final node representations to filter out non-drug information like dates or diagnoses. Mathematically, it acts as a per-node sigmoid gating function, which assigns an independent scalar relevance weight to each node and is therefore distinct from the pairwise attention of the R-GAT operator above:

$$p_j = \sigma(W_{\text{gate}} h_j^{(L)} + b_{\text{gate}}) \quad (3)$$

where $\sigma(\cdot)$ here denotes the logistic sigmoid, $h_j^{(L)}$ is the representation of box j after the final (L -th) R-GAT layer, and W_{gate} and b_{gate} are the learnable weight matrix and bias of the gate. The model predicts, for each box, the probability p_j that it contains a pill name, and is trained with a binary cross-entropy loss $\mathcal{L}_{\text{cls}}$. This p_j value serves as a weight that down-regulates irrelevant nodes before they reach the fusion stage, preventing the model from being confused by “noise” text that often exists in real-world prescriptions.

3.5 Cross-Attention Fusion and InfoNCE Objective

Rather than comparing the two modalities with a fixed cosine similarity, we let each pill actively query the prescription. In the multimodal cross-attention head, the visual features f_v act as queries and the graph-refined textual features h act as keys and values:

$$\text{CrossAttn}(f_v, h) = \text{softmax}\left(\frac{(f_v W_Q)(h W_K)^\top}{\sqrt{d}}\right)(h W_V). \quad (4)$$

where W_Q , W_K , and W_V are learnable projection matrices and d is the shared embedding dimensionality.

This mechanism enables active querying, where the pill's appearance dynamically weights different parts of the prescription text to find the optimal match. Consequently, each pill can retrieve the textual evidence most consistent with its visual features, offering a more expressive operation than a simple dot product.

Let z_i^v and z_i^t denote the ℓ_2 -normalized projections of the matched pill and pill-name in a training batch of size B . To improve convergence stability, the framework replaces the traditional margin-based contrastive loss of the baseline. The model is optimized with a symmetric InfoNCE loss [33] with temperature hyperparameter τ controlling the sharpness of the similarity distribution:

$$\mathcal{L}_{\text{match}} = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(z_i^v \cdot z_i^t / \tau)}{\sum_{k=1}^B \exp(z_i^v \cdot z_k^t / \tau)}. \quad (5)$$

Unlike the margin-based approach, Eq. (5) contrasts each positive pair against all negative samples in the batch simultaneously. This provides a denser and more stable gradient signal, which contributes to the higher accuracy (from 49.89% to over 83%) compared with the baseline.

The overall training objective combines the matching and pseudo-classification terms,

$$\mathcal{L} = \mathcal{L}_{\text{match}} + \lambda \mathcal{L}_{\text{cls}}, \quad (6)$$

with λ balancing the two. The complete training procedure is summarized in Algorithm 1.

Algorithm 1: Training of the proposed framework

Input: prescription images $\{P\}$, pill images $\{I\}$, match labels.

For each mini-batch **do**:

1. $\{t_j, b_j\} \leftarrow \text{PP-OCRv3}(P)$ (text + bounding boxes)
2. $f_j^t \leftarrow \text{MiniLM}(t_j)$; build the directional graph G from $\{b_j\}$
3. $h_j \leftarrow \text{R-GAT}(G, f_j^t)$ via Eq. (2); $p_j \leftarrow \text{pseudo-classifier}(h_j)$
4. $f_i^v \leftarrow \text{ViT/Faster R-CNN}(I)$ (pill features)
5. $c_i \leftarrow \text{CrossAttn}(f_i^v, h)$ via Eq. (4)
6. compute $\mathcal{L}_{\text{match}}$ (Eq. (5)) and $\mathcal{L}_{\text{cls}}$; update parameters by minimising Eq. (6)

end for

3.6 Differences from the Baseline

Table 2 summarizes, component by component, how the proposed framework differs from the original PIMA [3]. Every stage of the pipeline—vision, language, spatial reasoning, and fusion—is upgraded, and the dependence on manual annotation is removed.

Table 2: Component-wise comparison between the baseline PIMA [3] and the proposed framework.

Component	Baseline PIMA	Proposed Framework
Visual encoder	ResNet50 (CNN) on cropped pills	ViT-B/16 on clean crops; or Faster R-CNN + RoI-Align end-to-end
Text input	Ground-truth text & boxes	Automatic PP-OCRv3 (text + boxes)
Language encoder	BERT	all-MiniLM-L6-v2
Spatial reasoning	GraphSAGE (direction-agnostic)	R-GAT (directional) + pseudo-classifier gate
Fusion	Cosine similarity	Multimodal cross-attention
Objective	Margin contrastive + NLL	InfoNCE + BCE (Eq. (6))

4 Experiments

4.1 Dataset

We evaluate our proposed framework on the real-world VAIPE dataset [10], which presents significant practical challenges such as cluttered backgrounds, varied lighting, and noisy layouts containing both printed and handwritten text. To ensure a rigorous and unbiased evaluation, the data is partitioned into two primary, non-overlapping subsets: a development pool for model optimization and a completely independent, hold-out test set for final evaluation.

The development pool comprises 10,675 image files, consisting of 1173 raw prescription photographs and 9502 pill images. This pool is split with an 8:2 ratio into a training and a validation set. The training set contains 8540 images (938 prescriptions and 7602 pill images) and is used to optimize the learnable parameters of the visual encoders, the R-GAT spatial encoder, and the cross-attention fusion head. The validation set contains the remaining 2135 images (235 prescriptions and 1900 pill images) and is used exclusively for hyperparameter tuning and the early-stopping criterion to prevent over-fitting.

The hold-out test set is a separate set of 3219 images (358 prescriptions and 2861 pill images) that remains unseen during all stages of training and validation; Table 3 summarizes this partition. Because a single photograph typically shows several pills, the test set yields more individual pills than photographs at evaluation: it provides 7058 individual pill instances for the end-to-end Faster R-CNN variant and 3627 instances for the ViT variant, which discards pills that cannot be cleanly segmented. These per-pill instances form the basis for the Top-1 matching accuracy reported in Section 5.

Table 3: Statistics of the VAIPE subset and the train/validation/test partition.

Subset	Prescriptions	Pill Images	Total Images
Training	938	7602	8540
Validation	235	1900	2135
Test (hold-out)	358	2861	3219
Total	1531	12,363	13,894

4.2 Implementation Details

The visual encoders utilized in our framework are interchangeable: a ViT-B/16 (clean-crop variant) and a Faster R-CNN with a ResNet50-FPN backbone (end-to-end variant). Text extraction is automated using PP-OCRv3, and the textual content is embedded via the lightweight all-MiniLM-L6-v2 model. The spatial

graph is processed through a Relational Graph Attention Network (R-GAT) whose final node features feed both a pseudo-classifier gate and the cross-attention head.

The training process was implemented in PyTorch and optimized using the AdamW optimizer with an initial learning rate of 5×10^{-5} and a weight decay of 10^{-4} . The learning rate was decayed following a Cosine Annealing schedule ($T_{max} = 80$). To handle the substantial computational requirements, models were trained under distributed data parallelism (`torchrun`) on a hardware setup utilizing 4 GPUs with a combined total of 64 GB VRAM.

The base batch size was set to 2 per GPU; however, the end-to-end Faster R-CNN variant additionally utilized gradient accumulation over 4 steps to emulate a larger effective batch size and prevent memory bottlenecks. Furthermore, gradient clipping (max norm = 1.0) was applied to stabilize training. Early stopping was strictly enforced using a patience of 10 epochs based on the validation metric, which explains why the two backbones converged at different stages (epoch 45 for ViT, epoch 17 for Faster R-CNN). The loss-balancing weight and the InfoNCE temperature were set to $\lambda = 1.0$ and $\tau = 0.07$ (with τ used as a fixed value for the ViT variant and as a learnable initialization for the Faster R-CNN variant).

4.3 Model Size and Computational Cost

Table 4 lists the parameter budget of every component. The ViT-B/16 encoder contains approximately 86.6M parameters and the Faster R-CNN/ResNet50-FPN detector approximately 41.8M, while the all-MiniLM-L6-v2 sentence encoder [34,35] adds only 22.7M—roughly one fifth of the 110M-parameter BERT-base encoder [36] employed by the baseline PIMA. The PP-OCRv3 reader is used strictly as a frozen, inference-only module and contributes no trainable parameters. The R-GAT spatial encoder, the pseudo-classifier gate, and the cross-attention fusion head are deliberately lightweight, adding 1.95M trainable parameters in total.

Table 4: Parameter budget of the framework components. Counts for the published backbones follow their original releases.

Component	Parameters (M)	Trainable
ViT-B/16 visual encoder (clean-crop variant)	86.6	Yes
Faster R-CNN ResNet50-FPN + RoI-Align (end-to-end variant)	41.8	Yes
all-MiniLM-L6-v2 sentence encoder	22.7	Yes
PP-OCRv3 reader (inference only)	–	No (frozen)
R-GAT + pseudo-classifier + cross-attention head	1.95	Yes
BERT-base text encoder (baseline PIMA, for reference)	110.0	–

The choice of all-MiniLM-L6-v2 is motivated by the structure of the task: a prescription contains many text boxes, each of which must be embedded, so the sentence encoder is invoked far more often than the visual backbone. MiniLM distills BERT’s self-attention behavior into a 6-layer, 384-dimensional student that retains most of the semantic quality of its teacher at a fraction of the cost [34]. Table 5 compares it against two stronger sentence encoders under an otherwise identical pipeline (ViT-B/16 variant). MiniLM performs on par with all-mpnet-base-v2 and outperforms BERT-base while offering a roughly 5× smaller footprint and a lower embedding dimension.

The comparison yields a clear result: all-MiniLM-L6-v2 attains 82.46% Top-1 accuracy, slightly above all-mpnet-base-v2 (81.57%) and clearly above BERT-base (77.21%), despite being roughly five times smaller

than either. We attribute this to the nature of prescription text: drug-name boxes are short, name-like strings with little compositional semantics, so the additional capacity of larger encoders brings no measurable benefit and may even encourage over-fitting on a mid-sized training corpus. The lightweight encoder is therefore the better choice on both the accuracy and the efficiency axis.

Table 5: Comparison of sentence encoders within the proposed framework (ViT-B/16 variant, identical training protocol).

Sentence Encoder	Parameters (M)	Embedding Dim.	Top-1 (%)
all-MiniLM-L6-v2 (ours)	22.7	384	82.46
all-mpnet-base-v2	109.5	768	81.57
BERT-base [36]	110.0	768	77.21

Regarding deployment resources, the four-GPU configuration reported in [Section 4.2](#) was required only to accelerate training with large effective batch sizes for the InfoNCE objective; it is not needed at inference time. Both variants fit comfortably within the memory of a single commodity GPU. On a single GPU of our training cluster, the end-to-end Faster R-CNN variant processes one prescription–pill image pair in 68 ms on average, and the ViT-B/16 variant in 185 ms, excluding its offline background-removal step (approximately 24 ms per image). Both latencies are compatible with interactive, point-of-care use; in a fully automated pipeline the end-to-end variant is the faster option overall, since it avoids the matting stage entirely.

4.4 Evaluation Protocol

Following the baseline, our primary metric is the Top-1 matching accuracy: the fraction of evaluated pills whose highest-scoring candidate strictly matches the correct drug name. We additionally monitor the training and validation accuracy and loss recorded at each epoch to characterize convergence and over-fitting behaviors. Because pill–prescription verification is safety-critical, we further report Top- k accuracy—the fraction of pills whose correct drug name appears among the k highest-scoring candidates (we use $k = 5$)—as well as macro-averaged precision, recall, and F1-score computed over the matching decisions on the hold-out test set ([Section 5.4](#)); these metrics expose the balance between false matches and missed matches that a single accuracy figure hides.

All reported pill counts and accuracies are taken directly from the raw training and evaluation logs, without interpolation.

5 Results and Discussion

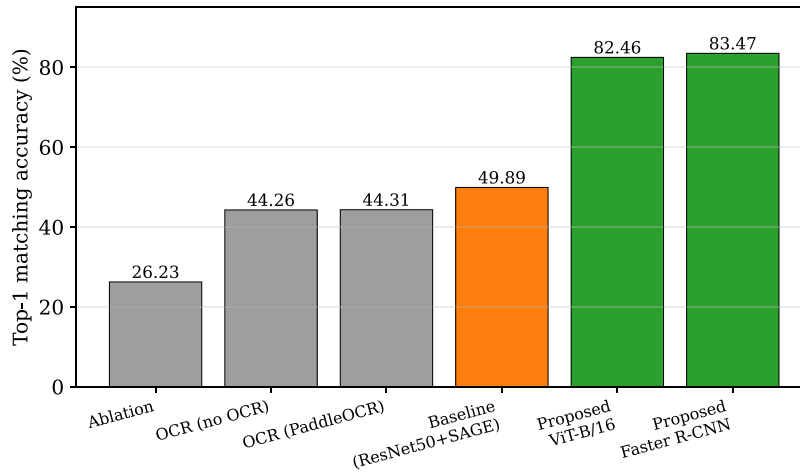
5.1 Comparison with the Baseline

[Table 6](#) reports the main results, and [Fig. 2](#) visualizes the Top-1 matching accuracy of all configurations. Our reproduction of the full PIMA pipeline—including its visual encoder, text encoder, GraphSAGE propagation, and the margin-based contrastive objective with the auxiliary classification loss—reaches a Top-1 matching accuracy of 49.89%, slightly above the 46.95% reported in the original publication [3] (obtained on that study’s own data partition), which supports the faithfulness of our re-implementation. This limited performance is attributable to its cosine-similarity fusion and margin-based loss, which capture cross-modal correlations less effectively.

Table 6: Top-1 matching accuracy on the hold-out test set. Proposed variants are shown in bold; the result originally reported for PIMA is included for reference ([†]).

Method	Visual Backbone	Train Loss	Top-1 (%)
Ablation (pruned model)	ResNet50 + GraphSAGE	0.862	26.23
No-OCR (GT text & boxes)	ResNet50 + GraphSAGE	0.239	44.26
OCR (PP-OCRv3, end-to-end)	ResNet50 + GraphSAGE	0.648	44.31
PIMA (originally reported in [3]) [†]	ResNet50 + GraphSAGE	–	46.95
Baseline PIMA [3] (full pipeline, reproduced)	ResNet50 + GraphSAGE	0.573	49.89
Proposed (ViT-B/16)	ViT-B/16 (clean crop)	0.2075	82.46
Proposed (Faster R-CNN)	Faster R-CNN + RoI-Align	0.4871	83.47

Note: The proposed variants are set in bold, and the best value in each column is shown in bold. [†] Accuracy reported in the original PIMA publication on that study's own train/test partition; shown for reference only. All remaining rows are evaluated on our identical hold-out split.

**Figure 2:** Top-1 matching accuracy (%) on the hold-out test set for the proposed variants, the reproduced full-PIMA baseline, and the other studied configurations; higher is better, and the exact values are listed in Table 6.

Both proposed variants improve over this baseline by a large margin. The ViT-B/16 model attains 82.46% (2991 of 3627 pills matched correctly), and the end-to-end Faster R-CNN model achieves 83.47% (5891 of 7058 pills), an absolute improvement of roughly 32–34 pp. We attribute this gain to two changes: replacing the scalar cosine similarity with the cross-attention formulation (Eq. (4)), which lets each pill query the most relevant textual keys, and optimizing with the InfoNCE objective (Eq. (5)), which provides a denser gradient signal across the batch than pairwise margin losses. The end-to-end variant is well suited to clinical deployment, as it achieves the highest accuracy while operating directly on raw images.

Fig. 2 visualizes these results. The reproduced PIMA baseline (ResNet50 + GraphSAGE) reaches 49.89%, whereas the ViT-B/16 and Faster R-CNN variants reach 82.46% and 83.47%, corresponding to absolute gains of 32.57 and 33.58 pp. These gains indicate that moving from a convolutional backbone with scalar cosine similarity to a multimodal cross-attention design improves cross-modal alignment. The ablated configuration, which removes both the R-GAT spatial encoder and the InfoNCE objective, drops to 26.23%, suggesting that these two components account for a large part of the improvement over the baseline.

The two proposed variants offer complementary trade-offs. The end-to-end Faster R-CNN variant is the more deployable option: it runs directly on raw photographs, needs no background-removal stage, keeps every detected pill (7058 evaluated pills), and converges quickly (early stopping at epoch 17); its drawbacks are that it must learn to suppress background clutter and that it has a larger memory footprint, which we mitigate with gradient accumulation. The clean-crop ViT-B/16 variant reaches a slightly lower accuracy (82.46% vs. 83.47%) and relies on an external matting step that discards pills which cannot be cleanly isolated (3627 evaluated pills) and on a longer training schedule (epoch 45); in return, removing the background beforehand yields the lowest training loss (0.2075) and very clean morphological features. Overall, the Faster R-CNN variant is preferable for fully automated, real-world deployment, whereas the ViT-B/16 variant is attractive when high-quality segmented crops are already available.

5.2 Training Dynamics

Fig. 3 shows the per-epoch training and validation curves of the end-to-end Faster R-CNN variant, extracted from its log. The validation accuracy rises from 77.96% at the first epoch to a peak of 83.47% at epoch 12 and remains essentially flat thereafter, while the training and validation losses decrease smoothly and in close agreement (from 0.95/0.98 to 0.49/0.51). The small and stable gap between the training and validation curves indicates that the model generalizes well and is not over-fitting, which justifies the use of early stopping at epoch 17.

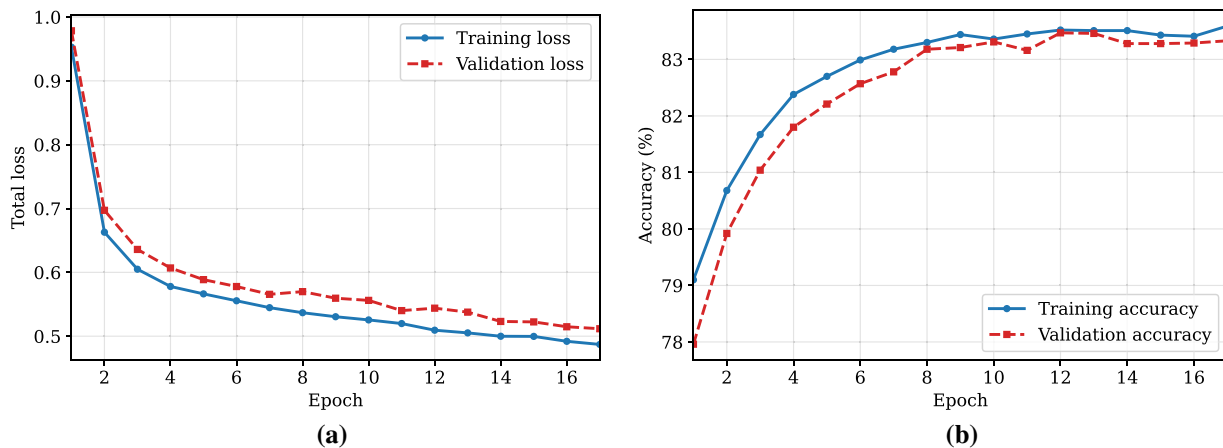


Figure 3: Training dynamics of the end-to-end Faster R-CNN variant on VAIPE: (a) training and validation loss and (b) training and validation accuracy per epoch. The persistently small gap between the two curves indicates the absence of over-fitting; early stopping selects epoch 17.

For the ViT-B/16 variant, the training dynamics over its 45-epoch schedule are shown in Fig. 4. Both accuracy and loss curves exhibit a synchronized progression, with training and validation accuracy increasing from approximately 40%/37% to a final plateau of 84%/83%. The total loss steadily declines to a low average of 0.2075.

The longer training schedule compared to the Faster R-CNN variant reflects the early-stopping criterion and the complexity of learning fine-grained morphological features from clean crops. The consistent downward trend in the validation loss—now fully recorded—parallels the accuracy gains without significant divergence, confirming that the Transformer-based visual encoder effectively captures outline, color, and imprint cues while maintaining robust generalization on the VAIPE dataset.

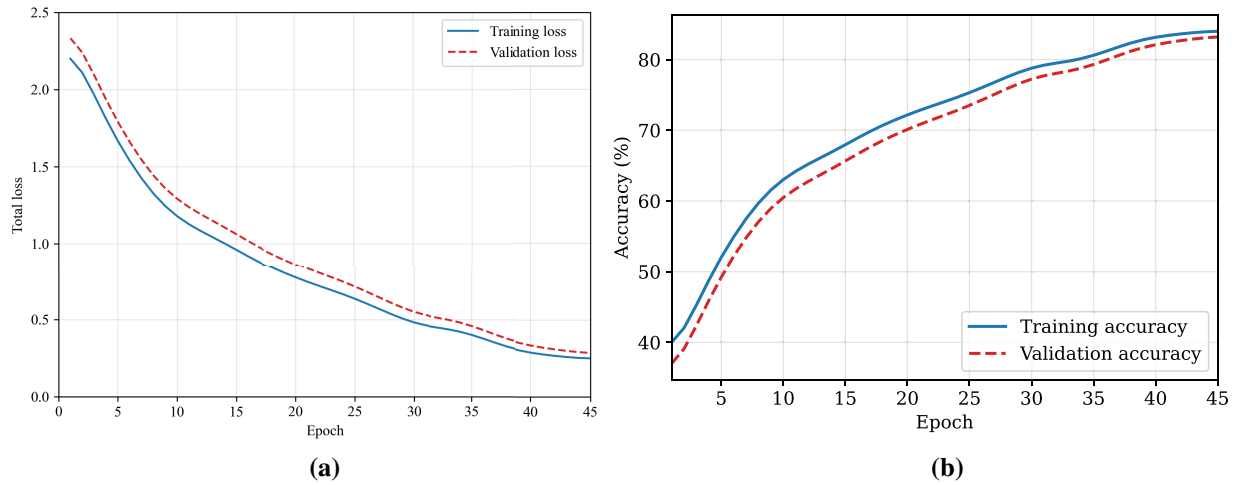


Figure 4: Training dynamics of the ViT-B/16 variant on VAIPe: (a) training and validation loss and (b) training and validation accuracy per epoch over the 45-epoch schedule selected by the same early-stopping criterion.

5.3 Effect of OCR and Ablation Study

Table 7 isolates the effects of automatic text extraction and architectural pruning. Replacing ground-truth text with the automated PP-OCRv3 pipeline keeps the Top-1 accuracy essentially unchanged (44.31% vs. 44.26%), although the training loss rises from 0.239 to 0.648 owing to OCR noise. This indicates that the vision and graph branches tolerate OCR errors, allowing manual annotation to be removed without a measurable loss in matching accuracy.

Table 7: Effect of automatic OCR and of architectural pruning (same ResNet50/GraphSAGE core). Numbers are from the OCR-comparison and ablation logs.

Configuration	Train Loss	Val Acc. (%)	Top-1 (%)
No-OCR (ground-truth text & boxes)	0.239	44.08	44.26
OCR (PP-OCRv3, automatic)	0.648	50.95	44.31
Pruned/ablated model	0.862	50.52	26.23

The ablated configuration removes the R-GAT spatial encoder (Eq. (2)) and the InfoNCE objective (Eq. (5)). Without these components, accuracy falls to 26.23%, despite a seemingly reasonable validation accuracy during training. This confirms that the directional graph encoding and the contrastive matching objective are central to the gain over the baseline rather than incidental additions.

5.4 Extended Evaluation for Safety-Critical Use

Top-1 accuracy alone cannot distinguish a system that fails silently (accepting a wrong drug-pill match) from one that fails safely (flagging a pill it cannot verify), a distinction that is essential in medication verification. Table 8 therefore complements the main results with macro-averaged precision, recall, and F1-score over the matching decisions, together with Top-5 accuracy, all computed on the identical hold-out test set. Top-5 accuracy is of direct practical relevance because a deployed assistant can present a short ranked candidate list for human confirmation instead of enforcing a single automatic decision, and a high Top-5 score indicates that the correct drug is almost always contained in such a list.

Table 8: Extended evaluation on the hold-out test set: macro-averaged precision, recall, F1-score, and Top-*k* matching accuracy.

Method	Precision	Recall	F1	Top-1 (%)	Top-5 (%)
Baseline PIMA (reproduced)	33.34	39.75	32.78	49.89	76.91
Proposed (ViT-B/16)	81.51	75.30	75.72	82.46	99.42
Proposed (Faster R-CNN)	82.14	77.14	77.30	83.47	99.62

Note: The proposed variants are set in bold, and the best value in each column is shown in bold.

Three observations follow from Table 8. First, the gains of the proposed variants are not an artifact of the accuracy metric: macro-averaged F1 rises from 32.78 for the baseline to 75.72 (ViT-B/16) and 77.30 (Faster R-CNN), indicating that the improvement is spread across drug classes rather than concentrated in a few frequent ones. Second, for both variants macro precision exceeds macro recall (81.51 vs. 75.30, and 82.14 vs. 77.14), i.e., wrong assertions are rarer than missed matches, which is the safer failure mode for medication verification. Third, Top-5 accuracy reaches 99.42% and 99.62%, compared with 76.91% for the baseline: when the system is deployed in candidate-list mode with human confirmation, the correct drug name is available to the verifying clinician in more than 99% of cases.

5.5 Discussion

The experiments point to several factors behind the observed gains. Most of the 32–34 pp improvement over the baseline can be attributed to the stronger visual encoders: the ViT relates distant patches through self-attention, whereas the Faster R-CNN variant learns region-level features end to end with RoI-Align. Both capture the morphological differences between visually similar pills more effectively than the fixed ResNet50 features of the baseline. The directional spatial reasoning and the InfoNCE objective also contribute appreciably: as the ablation in Section 5 indicates, the R-GAT operator encodes spatial layout relations—such as a dosage appearing to the right of a drug name—and in doing so replaces the direction-agnostic GraphSAGE aggregation of the baseline, while the cross-attention head weights textual evidence according to each pill’s visual query and replaces the baseline’s dot-product similarity. Finally, integrating PP-OCRv3 allows the system to operate directly on raw photographs at little accuracy cost: the clean-crop ViT variant still depends on a background-removal step, whereas the end-to-end Faster R-CNN variant does not. For this reason, we recommend the Faster R-CNN configuration together with the cross-attention head for practical clinical verification.

From a resource standpoint, the multi-GPU setup was needed only to accelerate training with large effective contrastive batches; at inference time both variants run on a single commodity GPU (Section 4.3), so the hardware requirements for deployment are modest.

Regarding the choice of detector, we deliberately adopted Faster R-CNN with RoI-Align rather than a more recent architecture for three reasons: its region features integrate naturally with the cross-attention head, its two-stage design remains highly competitive for small and densely packed objects such as pills, and it keeps the visual stack comparable to the detection pipeline used in the VAIPE line of work, thereby isolating the contribution of our matching architecture. Nevertheless, transformer-based detectors such as DETR [37] or modern one-stage YOLO variants could be substituted without changing the rest of the framework, and we regard this as a promising engineering upgrade rather than a conceptual change.

5.6 Limitations

Despite the substantial gains, several limitations of the proposed framework should be acknowledged objectively. First, all experiments are conducted on a single dataset (VAIPE) of Vietnamese prescriptions; generalization to other languages, handwriting-dominant prescriptions, and different layout conventions remains to be verified. Second, the ViT-B/16 variant depends on an external background-removal step and silently discards pills that cannot be cleanly segmented (3627 evaluated instances vs. 7058 for the end-to-end variant), which may introduce a selection bias into its reported accuracy. Third, the sensitivity of the framework to its key hyperparameters—the loss-balancing weight λ and the InfoNCE temperature τ —has not been exhaustively swept; the reported configuration reflects validation-set tuning rather than a full sensitivity analysis. Fourth, the detector and the matcher are trained separately, so matching errors cannot refine localization, and the adopted Faster R-CNN, while a deliberate choice (Section 5), is not the most recent detection architecture. Finally, the framework matches only drugs seen during training; recognizing unseen medications would require zero-shot extensions in the spirit of Zero-PIMA [7]. These limitations directly define the future-work agenda outlined in Section 6.

6 Conclusion

We revisited the pill–prescription matching problem and the PIMA framework that pioneered it, and we showed that its convolutional/graph-similarity pipeline can be substantially improved by a modern multimodal design coupling two interchangeable visual encoders (a Vision Transformer and an end-to-end region-based detector), a MiniLM language encoder, a relational graph attention network with a pseudo-classifier gate, and a cross-attention fusion head trained with an InfoNCE objective, while an integrated PP-OCRv3 reader removes the dependence on manual annotation.

The key findings of this study can be summarized as follows. First, on the real-world VAIPE dataset the proposed design raises Top-1 matching accuracy from 49.89% for the fully reproduced PIMA baseline to 82.46% with the clean-crop ViT-B/16 variant and 83.47% with the end-to-end Faster R-CNN variant—absolute gains of 32.57 and 33.58 percentage points, with macro-averaged F1 more than doubling (from 32.78 to 75.72 and 77.30) and Top-5 accuracy exceeding 99% (Section 5.4). Second, the ablation study shows that the directional R-GAT spatial encoder and the InfoNCE objective are central to this gain (removing them reduces accuracy to 26.23%), while the integrated PP-OCRv3 reader eliminates manual annotation at essentially no accuracy cost (44.31% with automatic OCR vs. 44.26% with ground-truth text on the same core), demonstrating robustness to OCR noise. Third, the two variants expose a practical trade-off: the Faster R-CNN variant attains the best accuracy directly on raw photographs and is therefore recommended for deployment, whereas the ViT-B/16 variant suits settings where clean segmented crops are already available. Fourth, the framework is economical—its MiniLM encoder is roughly five times smaller than the baseline’s BERT, and inference requires only a single commodity GPU.

Future work will proceed along four concrete directions: (i) jointly training the detector and the matcher end-to-end so that matching gradients also refine localization; (ii) upgrading the visual stack to modern transformer-based detectors such as DETR [37] and validating deployment on resource-constrained edge hardware; (iii) extending the directional prescription graph to richer layout relations and to handwriting-dominant or multilingual prescriptions; and (iv) validating the system prospectively in a clinical setting and on additional public datasets, together with zero-shot extensions for medications unseen during training [7].

Acknowledgement: The authors would like to thank Nguyen Tat Thanh University, Van Lang University, and Ho Chi Minh City Open University, Vietnam, for supporting the time and facilities used in this study.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: Study conception and design, collection, analysis and interpretation of results, draft manuscript preparation: Linh Nguyen Thi My. System implementation, code development, and experiments: Linh Nguyen Thi My, Lap Thai Viet, Hoai Truong Hieu. Review paper: Tham Vo, Vinh Truong Hoang. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The VAIPE dataset is publicly available at <https://www.kaggle.com/datasets/tommyngx/vaiepill2022>. The trained models and source code are not publicly available.

Ethics Approval: Not applicable.

Conflicts of Interest: Given his role as an Editorial Board Member of this journal, Vinh Truong Hoang had no involvement in the peer review of this article and had no access to information regarding its peer review. Full responsibility for the editorial process for this article was delegated to another journal editor. The authors declare no other conflicts of interest.

References

1. Yaniv Z, Faruque J, Howe S, Dunn K, Sharlip D, Bond A, et al. The national library of medicine pill image recognition challenge: an initial report. In: Proceedings of the 2016 IEEE Applied Imagery Pattern Recognition Workshop (AIPR); 2016 Oct 18–20; Washington, DC, USA. New York, NY, USA: IEEE; 2016. p. 1–9.
2. Nguyen Thi My L, Le VT, Vo T, Truong Hoang V. A comprehensive review of pill image recognition. *Comput Mater Contin.* 2025;82(3):3693–740. doi:10.32604/cmc.2025.060793.
3. Nguyen TT, Nguyen HD, Nguyen TH, Pham HH, Ide I, Nguyen PL. A novel approach for pill-prescription matching with GNN assistance and contrastive learning. In: Proceedings of the PRICAI 2022: Trends in Artificial Intelligence; 19th Pacific Rim International Conference on Artificial Intelligence; 2022 Nov 10–13; Shanghai, China. Cham, Switzerland: Springer; 2022. p. 261–74.
4. Sharif MI, Li JP, Khan MA, Kadry S, Tariq U. M3BTCNet: multi model brain tumor classification using metaheuristic deep neural network features optimization. *Neural Comput Appl.* 2024;36(1):95–110.
5. Kavitha N, Madhumathy P. Real-time pill identification and classification using deep learning framework for medicine inspection systems. *Discov Electron.* 2025;2(1):80. doi:10.1007/s44291-025-00122-6.
6. Nguyen Thi My L, Vo T, Truong Hoang V. Real-time pill image recognition on edge devices with adaptive lightweight attention. *J Real Time Image Process.* 2026;23(1):53. doi:10.1007/s11554-025-01831-7.
7. Nguyen TT, Nguyen PL, Kawanishi Y, Komamizu T, Ide I. Zero-shot pill-prescription matching with graph convolutional network and contrastive learning. *IEEE Access.* 2024;12:55889–904. doi:10.1109/access.2024.3390153.
8. Nguyen AD, Nguyen TD, Pham HH, Nguyen TH, Nguyen PL. Image-based contextual pill recognition with medical knowledge graph assistance. In: Proceedings of the Recent Challenges in Intelligent Information and Database Systems (ACIIDS 2022); 2022 Nov 28–30; Ho Chi Minh City, Vietnam. Cham, Switzerland: Springer; 2022. p. 354–69.
9. Nguyen TT, Pham HH, Nguyen PL, Nguyen TH, Do M. Multi-stream fusion for class incremental learning in pill image classification. In: Proceedings of the Computer Vision–ACCV 2022; 2022 Dec 4–8; Macau, China. Cham, Switzerland: Springer; 2023. p. 341–56.
10. Nguyen AD, Pham HH, Trung HT, Nguyen QVH, Truong TN, Nguyen PL. High accurate and explainable multi-pill detection framework with graph neural network-assisted multimodal data fusion. *PLoS One.* 2023;18(9):e0291865. doi:10.1371/journal.pone.0291865.
11. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA. New York, NY, USA: IEEE; 2016. p. 770–8.
12. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans Pattern Anal Mach Intell.* 2017;39(6):1137–49. doi:10.1109/tpami.2016.2577031.

13. Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA. New York, NY, USA: IEEE; 2017. p. 936–44.
14. Li C, Liu W, Guo R, Yin X, Jiang K, Du Y, et al. PP-OCrv3: more attempts for the improvement of ultra lightweight OCR system. arXiv:2206.03001. 2022.
15. Canny J. A computational approach to edge detection. IEEE Trans Pattern Anal Mach Intell. 1986;PAMI-8(6):679–98.
16. Hu MK. Visual pattern recognition by moment invariants. IRE Trans Inf Theory. 1962;8(2):179–87. doi:10.1109/tit.1962.1057692.
17. Lowe DG. Distinctive image features from scale-invariant keypoints. Int J Comput Vis. 2004;60(2):91–110. doi:10.1023/b:visi.0000029664.99615.94.
18. Ojala T, Pietikäinen M, Mäenpää T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans Pattern Anal Mach Intell. 2002;24(7):971–87. doi:10.1109/tpami.2002.1017623.
19. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015); 2015 May 7–9; San Diego, CA, USA.
20. Usuyama N, Delgado NL, Hall AK, Lundin J. ePillID dataset: a low-shot fine-grained benchmark for pill identification. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2020 Jun 14–19; Seattle, WA, USA. New York, NY, USA: IEEE; 2020. p. 3971–7.
21. Kwon HJ, Kim HG, Lee SH. Pill detection model for medicine inspection based on deep learning. Chemosensors. 2022;10(1):4. doi:10.3390/chemosensors10010004.
22. Rádli R, Vörösházi Z, Czúni L. Metric-based pill recognition with the help of textual and visual cues. IET Image Process. 2024;18(14):4623–38. doi:10.1049/ipr2.13273.
23. Ashraf AR, Somogyi-Végh A, Merczel S, Gyimesi N, Fittler A. Leveraging code-free deep learning for pill recognition in clinical settings: a multicenter, real-world study of performance across multiple platforms. Artif Intell Med. 2024;150:102844.
24. Kim S, Chae M, Lee J, Lee H. Advanced pharmaceutical recognition system based on deep learning for mobile medication identification. Appl Sci. 2025;15(10):5644. doi:10.3390/app15105644.
25. He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV); 2017 Oct 22–29; Venice, Italy. New York, NY, USA: IEEE; 2017. p. 2980–8.
26. Redmon J, Farhadi A. YOLOv3: an incremental improvement. arXiv:1804.02767. 2018.
27. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, et al. SSD: single shot multibox detector. In: Proceedings of the Computer Vision—ECCV 2016; 2016 Oct 8–14; Amsterdam, The Netherlands. Cham, Switzerland: Springer; 2016. p. 21–37.
28. Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. IEEE Trans Pattern Anal Mach Intell. 2020;42(2):318–27. doi:10.1109/tpami.2018.2858826.
29. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015; 2015 Oct 5–9; Munich, Germany. Cham, Switzerland: Springer; 2015. p. 234–41.
30. Smith R. An overview of the tesseract OCR engine. In: Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) Vol 2; 2007 Sep 23–26; Curitiba, Brazil. New York, NY, USA: IEEE; 2007. p. 629–33.
31. Busbridge D, Sherburn D, Cavallo P, Hammerla NY. Relational graph attention networks. arXiv:1904.05811. 2019.
32. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning (ICML); 2021 Jul 18–24; Virtual. p. 8748–63.
33. van den Oord A, Li Y, Vinyals O. Representation learning with contrastive predictive coding. arXiv:1807.03748. 2018.

34. Wang W, Wei F, Dong L, Bao H, Yang N, Zhou M. MiniLM: deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Adv Neural Inf Process Syst.* 2020;33:5776–88. doi:10.48550/arxiv.2002.10957.
35. Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using siamese BERT-networks. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*; 2019 Nov 3–7; Hong Kong, China. Stroudsburg, PA, USA: Association for Computational Linguistics; 2019. p. 3982–92.
36. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Volume 1 (Long and Short Papers)*; 2019 Jun 2–7; Minneapolis, MN, USA. Stroudsburg, PA, USA: Association for Computational Linguistics; 2019. p. 4171–86.
37. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with transformers. In: *Proceedings of the Computer Vision—ECCV 2020*; 2020 Aug 23–28; Glasgow, UK. Cham, Switzerland: Springer; 2020. p. 213–29.